DOI : 10.6151/CERQ.2016.2401.04

《研究紀要》

校務研究資料庫的建構與分析應用

曾元顯

摘　要

本文根據教育科學、圖書館學、資訊工程等領域的知識，結合多年來處理校內外資料庫的實務經驗，以及對近年技術產品的瞭解，闡述校務研究資料庫建置的較佳實施概念，並分析比較各種建置方案的適用時機與優缺點，最後以三項代表性的具體實例說明本文提及之實施概念的綜效。

具體而言，本文介紹採、編、典、藏、用五項可持久運作的資料庫建構作業流程與注意事項，說明資料正規化與反正規化的用處，透過概念驗證作業提供與國內廠商互動的經驗，並以實際校務數據的分析案例，分享應用經驗。

整體而言，資料蒐整作業（即：採、編、典、藏）仍是最費時、費力的流程，一旦完成，後續的分析運用便容易進行。目前的視覺化分析工具將越來越便利，讓各類型使用者得以更有效率的從大量資料中發現特殊樣態、形成假說，進而對資料做各種查詢與探索，以獲得具體事證支持決策。展望未來，除了視覺化工具越受依賴外，事件演進模擬技術，也將扮演重要角色，其可讓使用者事先知道各種因素變化後的最終結果，讓分析平台更具價值。

關鍵詞：資料一致性、資料正規化、資料素養、資料倉儲、視覺化分析

曾元顯（通訊作者），國立臺灣師範大學研究員

電子郵件：samtseng@ntnu.edu.tw

投稿日期：2015 年 08 月 03 日；修正日期：2015 年 10 月 06 日；接受日期：2015 年 12 月 31 日

March, 2016, Vol. 24, No. 1, pp. 107-134
DOI : 10.6151/CERQ.2016.2401.04

《Technical Note》

DEVELOPMENT AND APPLICATION OF DATABASES FOR INSTITUTIONAL RESEARCH AND ANALYSIS

Yuen-Hsien Tseng

ABSTRACT

This article elaborates on the possible best practice of developing databases for institutional research and analysis, based on the knowledge of Educational Science, Library Science, and Information Engineering, years of experience in developing educational databases, and a recent survey of related technology and products. Several developing options are compared to show their benefits and disadvantages under different conditions. Three representative analysis tasks are reported to verify and show the synergy of the mentioned ideas and experience.

In particular, this article proposes a sustainable workflow: (1) data collection and aggregation, (2) cataloguing, (3) regulation, (4) archiving, and (5) usage, and describes their must-known caveats. The application situations of data normalization and de-normalization are described. Capability of domestic vendors of related products is briefly mentioned based on a proof-of-concept testing. And finally, real-world institutional analyses are conducted to share our experience.

Overall, the first four processes in the above workflow are most time-consuming and costly. Once data have been well prepared, recent visualization analysis tools allow users to easily discover meaningful patterns and inspire hypotheses, and allow them to explore the database to find evidence to support their hypotheses and decisions. In the future, we expect that event evolution simulation techniques, which allow users to foresee the results given various input scenarios, could play an important role in educational data analysis, in addition to the maturing data visualization tools.

Keywords: *data consistency, data normalization, data literacy, data warehouse, visualization analysis*

Yuen-Hsien Tseng (corresponding author), Research Fellow, National Taiwan Normal University.

E-mail: samtseng@ntnu.edu.tw

Manuscript received: August 3, 2015; Modified: October 6, 2015; Accepted: December 31, 2015

壹、前言

一、校務資料庫的重要性

校務資料庫不僅是累積學校辦學的知識庫，也是支援學校日常運作、決策未來的重要資產。相對於中小學而言，大學的校務發展，有較高的自主性。此自主性伴隨而來的責任，是大學（或同類的高等教育機構）需要在全球化浪潮的影響下，與世界其他大學競爭傑出的教研人員以及優秀學生，才能讓大學本身蓬勃發展，培育國家及世界未來的棟梁。這意味著，大學需要投入校務資料庫的建置、發展，蒐集並利用各種數據，進行分析，做為決策的參考依據、方案實施的成效評估，進而累積、活化與善用組織知識（彭森明，2013；Hossler, Kuh, & Olsen, 2001; Nonaka, 1994）。

舉例而言，美國第一位非裔（黑人）總統當選後，夏威夷的奧阿厚學院，即今日普納荷（Punahou）中學，變成世界新聞的焦點[1]，一夕之間這間中學成為世界上唯一有中國與美國領導人都曾就學過的學校（孫中山 1883 年曾於該校就讀，後為中華民國臨時大總統；歐巴馬於 1979 年畢業，為美國第四十四任總統）[2]。此種獨特性不僅提高該校的知名度，對其後的招生與續存，都會產生正面效益。然而，要有這樣的巧合，其學籍資料必須保存 96 年，才有證據提供報導。而臺師大的國語教學中心，自 1956 年成立起，因歷史因素，為全世界學習華語的重要機構。例如，前澳洲總理陸克文、前日本首相橋本龍太郎皆曾在國語中心就學過，其意義跟 Punahou 中學有中、美兩國的領導人曾就學過類似。因此，國語中心仍是許多外國人學習華語的首選，也是政府單位極為重視的華語推廣單位。

二、校務資料庫的範圍與整合

校務資料涵蓋的範圍廣泛，包括：學生、校友、家長、雇主、教務、學務、教職員、研究、經費、行政、各項資源（如空間、設備、水電、網路用量）等跟學校的人、事、時、地、物有關的資料，都需要蒐集、整理、保存成資料庫。就內容特性而言，這些資料可分為：客觀資料與主觀資料，

[1] 以「歐巴馬、孫中山、夏威夷」搜尋 Google 在 2011 年 1 月 1 日至 2011 年 12 月 31 日的網頁，可看到多種報紙、部落格報導孫中山為歐巴馬的學長此件趣聞。
[2] 參見維基百科：http://en.wikipedia.org/wiki/Punahou_School，存取日期 2015 年 5 月 24 日。

前者如：個人資訊、校方記錄、外部官方數據等事實性資料；後者如：意見調查、評估建議等觀感、印象型態的資料。就記錄的方式，這些資料可分為：結構化資料與非結構化資料，前者即每筆資料皆有相同欄位、屬性的資料，如：學生修課與成績記錄、學習平台存取記錄、研究著作目錄等；後者多為以自由文字描述的內容，如：評估意見、學習心得、研究報告等，甚至是超越文字的圖片、動畫、影音檔案等內容。

從上可知，校務資料並非新事物，大多數資料，大學本身平時已有蒐集、整理與運用。而根據校務資料進行的統計與分析亦非新作為，有些報表平時已在產製，提供行政上的運用。然而，個別系統的分析報表，若無特異之處，經常僅僅成為例行業務的產出紀錄。當報表數據異常時，才會引起注意而進行深入的分析探究。特別是結合各種內外在環境資訊所進行的研究，常成為解答數據異常的關鍵作為。

這幾年，臺灣各大學和研究所招生不足的嚴重程度，單靠教務資訊系統，難以預測與掌握真正原因。若每年皆招滿人數的話，不會刻意討論這些招生數據。一旦招生數量大幅波動，會立即引起整間學校的關注。不僅學校經費、資源受到影響，各學院院長和系所主管，甚至連教師本身，也會成為利害關係人，為此必須投入各種關注與研究來處理這項危機。而此種現象，若結合外部人口出生率的資料與各產業的就業率分布，將可預測並解釋此項數據的波動，從而預先做出因應方案。

由於校務資料涵蓋的範圍廣泛，且各單位在建置相關資料庫時，皆以完成自身任務為主，數十年下來，校務資料實際上分散在各行政或學術單位。在進行校務研究或教育資料分析時，需事先全面盤點各單位系統，以瞭解並掌握各項校務資料，並在符合個人資料保護法、政府資訊公開法的情況下，制訂資料整合的作業程序，以便對整體的校務資料做妥善的運用，發揮其應有的價值。

三、教育大數據

一旦各類校務資料交相作用，其所產生出來的數量（Volume）、成長速度（Velocity）與多樣性（Variety），將符合近年來常被提及的大數據（Big Data）定義與重要性。雖然一般所稱的大數據，是指無法以傳統方式在合理的時間內分析處理完上述特性的資料，但也有人持不同的看法

（李欣宜，2015）：「他們以為大數據就是指大數目的數據，事實上……，我們真正在尋找的是非傳統的、而且未曾被挖掘過的資料，並且從這些資料中去提煉出價值，我相信在五年內我們就不會再使用「大數據」這個詞了，到頭來大數據就只是資料而已……」、「一般來說我們用 3V 定義大數據，數量（Volume）、速度（Velocity）與多樣性（Variety），其中我認為最重要的是多樣性，資料不只來自那些傳統管道，有更多來自非傳統管道的非傳統資料產生，我認為價值（Value）是第四個 V，人們常常忘了這件事，他們專注於技術，卻忘了創造價值，但這卻是一個大數據計畫能否成功的關鍵：這不只關乎技術，而是你能用技術創造出什麼價值。」

換言之，基於校務資料所做的校務研究，或是教育大數據分析，其重點在於能否發現「好的議題」、「有用的資料來源」與「適當的研究方法」，以創造出學校所需的分析結果（Rios-Aguilar, 2015）。

一般而言，校務資料經常需要配合外部資料一起做交叉分析，才會變得更有價值。因此，除了常見的敘述統計用來初窺資料庫內容的樣貌，結合外部資料的深度探索（如：畢業生流向、雇主意見、學校聲望、教師論文被引用情況），或是比較性分析（與國內外規模、領域相當的機構進行對照與反思），常能讓校務資料庫，發揮更大的效益。

本文根據教育科學、圖書館學、資訊工程等領域的知識，結合作者多年來處理校內外資料庫的實務經驗，以及對近年技術產品的瞭解，闡述校務研究數據資料庫建置的較佳實施概念，並分析比較各種建置方案的適用時機與優缺點，將散落各處的隱性（implicit）、默會（tacit）知識外顯化，最後以三項代表性的具體實例說明校務數據的分析與應用。

貳、校務研究數據資料庫的建置概念

如前所述，大學本身早已建置各項校務資料的資料庫，以進行日常的資料蒐集、處理與運用。本文所指的校務研究資料庫，是指圖 1 中的「數據分析平台」（Data Analytics Platform），用以整合大學內、外部的各種資料，並進行綜合分析與應用的大型數據庫。

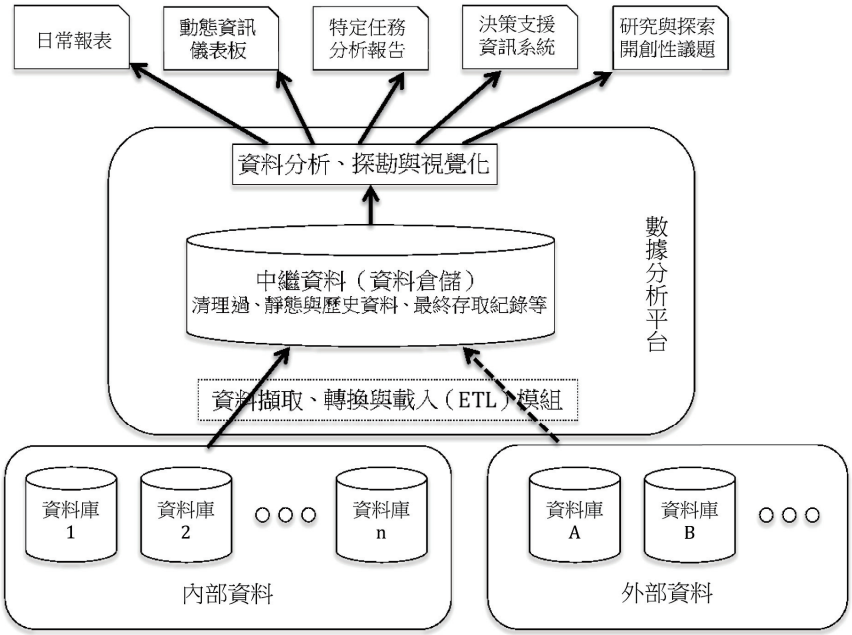


圖 1 校務研究資料庫架構。低層的資料庫，可涵蓋：結構化資料（如：學生基本資料、教學平台存取紀錄等），半結構化資料（如：開課資料與其課程大綱、意見調查的封閉與開放問卷內容等），以及非結構化資料（如教材內容、文件檔案等）。

在圖 1 的數據分析平台中，包含三部分：一是將內外部各類資料進行擷取、轉換、載入（Extract, Transform, and Load, ETL）的軟體模組；二是儲存大量資料的資料倉儲（Data Warehouse）軟硬體；三是提取資料以進行各種分析、探勘與視覺化的軟體工具。其中資料倉儲與 ETL 的理論與技術發展，已有 20 多年的歷史（Inmon, 1992），特別是在商業環境，相關產品與應用已經相當成熟（詹文男、羅瑋君，2001）。但近年來因應大量數據的出現與分析需求，資料倉儲的概念與技術，產生很大的轉變：不僅既有的系統，如目前使用廣泛的關聯式資料庫（Relational Database）（Codd, 1970），在效能上有大幅的改善；新的系統，如 Hadoop 平台（Borthakur, 2007），更是跳出傳統的框架，對特定的巨量資料有極佳的處理效能。至於資料分析（Rice, 2006）、探勘（Han, Kamber, & Pei, 2011; Larose, 2014）與視覺化（Ware, 2012）部分，則是持續不斷演進的技術與產品（Plaisant, 2004）。

數據分析平台在軟、硬體方面，近幾年雖然有大幅的進步與改善，然其有效的建立與良好的運用，實務上還需顧慮多種面向的議題。在建構校務研究資料庫的分析應用時，針對這些議題，可藉助圖書館學與資訊工程學的知識與經驗，以「採、編、典、藏、用」五項口訣式的概念，化繁為簡的建立資料處理的正確觀念，並綜整、摘要其對應的解決方案與注意事項（曾元顯，2014）。此五項概念的細節描述如下。

一、「採」即資料「採訪」之意，是指針對有潛在價值的資料進行界定與採集。

這項問題看似簡單，但極為重要，否則垃圾進、垃圾出，失去分析的意義，並浪費資料蒐集與處理的各種成本（人力、時間、儲存、計算等）。

另外，這項工作需要教育專業領域知識，才能了解哪些資料對校務研究以及教育大數據具有目前的重要性、未來的前景性、以及橫向的互用性（inter-operability）。舉例而言，國外很多大學新生入學時，要填寫的個人相關資料，多到常需數小時方能完成；而臺師大的新生入學時，也需填寫約 70 項的欄位內容。這麼多欄位看似繁瑣，其實是經過教務處、學務處多番討論後訂定的最少所需資料。

事實上，對於運作良好的大學來說，各單位通常已有多種資料庫系統支援日常行政運作，此時可透過資料庫盤點作業，全面清查學校的資料資源，做為建立資料倉儲的依據。以臺師大為例，各行政單位約 240 個資料庫系統，總共約 3,800 張資料表、52,000 個欄位。經過進一步的比對、篩選，將依照各類議題分析的進度，逐步將這些欄位資料載入到資料倉儲中。

此外，若無法以任何管道蒐集到的資料，而需要進行個人訪談或問卷調查時，也要具備訪談技巧、問卷設計能力、抽樣施測技術（Simone, Campbell, & Newhart, 2012），且對提高問卷回收機制，排除隨意回應、無效的問卷，也需具備相關經驗（Dey, 1997）。

二、「編」在圖書館學意為資料「編目」（編列目錄）、資訊組織之意，在資訊工程學，意指資料庫的設計。不管那個領域，其意義在對蒐集來的資料，進行嚴謹的詮釋資料（meta-data）描述，以及資料的主題分析、分門別類的組織，以方便後續的應用。

針對結構化的資料，可運用圖書館學的權威控制（authority control）概念與關聯式資料庫的設計理論與工具，做到如下的效益：

甲、確保資料蒐集時的一致性：使用者輸入或匯入資料時立即檢查一致性，確實做好權威控制（Tillett, 2004），亦即讓相同意義的欄位，都有一致的內容。例如，在調查大學新生家庭背景的父母親職業，或是畢業生的職業時，可參考「中華民國職業標準分類[3]」，而不是讓使用者隨意輸入或自訂職業名稱，造成後續分析的困擾。該職業表分成 10 大類、39 中類、125 小類、380 細類，其以聯合國國際職業標準分類為基本架構，並兼顧我國國情，讓各機關的資料統一，有利於互相的連結分析。而從不同來源蒐集類似的資料進行整合時，也應盡可能自動處理與比對，做好資料清理（data cleansing）（Müller & Freytag, 2003）與正確的紀錄連結（record linkage[4]）（Fellegi & Sunter, 1969; Jin, Li, & Mehrotra, 2003）。

乙、資料庫正規化以降低欄位內容的重複性，並讓後續新增欄位更具彈性：資料庫正規化是資料庫設計的核心理論與實務，是資訊相關系所獨特的重點課程。如表 1 所示，以選課資料表為例，若同時記載學生與課程的詳細資料於一張資料表，則一位學生修二門課時，學生的詳細資料（如：姓名、年級）在該資料表上就會重複二次；若一門課有五十位同學選修時，則該門課的詳細資料（如：上課時間、地點）就會重複五十次。資料表正規化後，此選課資料表會分成三個資料表儲存個別的資料，如圖 2 所示：學生資料表、課程資料表以及選課資料表，最後的選課資料表只需紀錄學生學號、課程代號即可。而後續若需增加課程項目欄位，如授課教師的姓名，只要在課程資料表增加即可，如圖 3 所示，而不用像表 1 那樣重複紀錄五十次。資料庫正規化不僅大幅降低資料庫重複儲存相同的欄位內容，新增欄位項目也更具彈性，同時也可降低人員重複輸入相同欄位內容時，造成資料不一致的問題（如：誤觸鍵盤造成的漏字、冗詞，或是輸入錯字、名稱前後不一致等現象）。

[3] http://www.stat.gov.tw/public/Attachment/141413555071.pdf，存取日期 2015 年 6 月 6 日。
[4] 範例與作法，詳見：https://en.wikipedia.org/wiki/Record_linkage，存取日期 2015 年 7 月 13 日。

表 1　選課資料表

CourseName	CourseTime	CoursePlace	SName	SGrade
電腦概論	星期二第二節	誠 101 教室	陳○○	一
電腦概論	星期二第二節	誠 101 教室	李○○	一
程式設計	星期四第五節	教 403 教室	陳○○	一
程式設計	星期四第五節	教 403 教室	蔡○○	二
程式設計	星期四第五節	教 403 教室	李○○	一

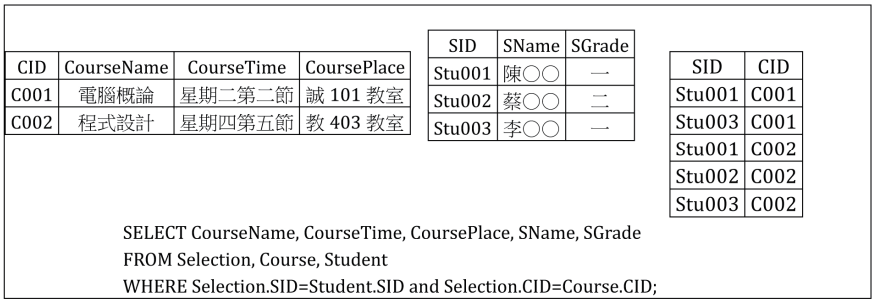


圖 2　正規化後的選課資料表。資料表名稱分別為：Course、Student、Selection，透過圖下面的 SQL 語法的 SELECT 命令，可將三個資料表反正規化，還原成表 1 的結果

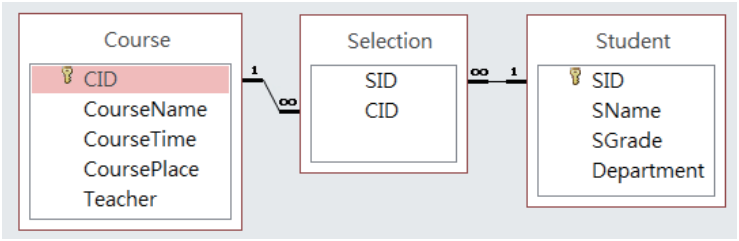


圖 3　選課資料表關聯圖

丙、建立資料表關聯圖，呈現資料之間的關係，方便後續的應用：上述資料表正規化後，資料項目分散在各個表中，不易得知彼此之間的關係，透過資料表的主鍵（primary key，如圖 2 或圖 3 中 Cid 為資料表 Course 的主鍵）與外鍵（foreign key，如圖 3 中 Cid 與 Sid 都是資料表 Selection 的外鍵）的定義，可建立資料表之間的關係，並視覺化呈現，不僅讓人輕易了解資料項目之間的關係，更能便利各種臨時、非預期的查詢與運用。如圖 2 的例子原本分散的多個資料表，透過 SQL（Structured Query Language）查詢語言的 SELECT 命令，可恢復原先的選課詳細資料表（亦即，將其反正規化，可送到資料倉儲），更可提供個別學生的選課表（供其上學使用），或是給老師修課名單（供其點名計分用）。

針對非結構化的資料，則可運用文字探勘（Text Mining）技術，如：關鍵詞擷取、關聯詞分析、內容摘要、事件歸類、主題分類、模糊比對、語意查詢等，以自動或半自動化的方式將其轉換成結構化資料、進行權威控制，或是對其直接進行查詢、探索與運用（曾元顯，2014；Feldman & Sanger, 2006; Tseng, Lin, C.-J., & Lin, Y.-I., 2007）。

三、「典」即「法典」，意指建立資料蒐整的標準作業流程與使用規範。

校務研究的數據資料，在蒐集、處理、利用、分享、管理等各個流程，不僅要兼顧法律面的「個人資料保護法」與「政府資訊公開法」、訂定適合機構本身以及分析目標的規範、讓資料擁有者了解其資料被使用的狀況（使其樂意提供後續資料），也要將資料蒐集的策略與過程、欄位項目的細節與變遷，詳加記載說明，使後續的分析者知道資料的來龍去脈，做出正確的解讀與運用。

特別是在資料傳承、分享與整合時，資料字典（Data Dictionary）（Zhang, Wang, & Han, 2011; Kuhn, 2013; Narayan, 1988）、資料庫設計等規範性文件特別重要。我們在協助整合全國教育相關資料庫以提供進階的應用時，便發現主要問題之一是缺乏資料項目的精確定義、演變過程、蒐集與處理的詳細描述與流程規範，以至於整合時所需成本極高，其對應的資料倉儲建構困難。

四、「藏」即資料「儲藏」之意。

資料不僅只是放置到儲存設備與系統，做到備份（上線的資料損毀時，可從備份資料復原回來）與備援（上線的系統失效時，可從備援系統即時接手運作）等機制（鍾沛原、曾賢寶、楊嘉麗、李柏毅、蔡一郎，2014），從前述的各種處理到資料儲藏，這整個流程，應當要盡可能自動化，而不要仰賴手工處理例外資料，以便讓儲藏的資料具備可重製性。

特別是各種衍生性資料（從原始資料整理、推演出的資料項目），需自動化處理，且整理、衍生規則要說明清楚，任何例外處理，需詳加記載，並以程式自動驗證，使蒐整的儲藏資料，具備可從原始資料自動重製，而能得到相同儲藏資料的特性。

我們的經驗顯示，原始資料常會有各種原因產生變動（如：部分資料提供單位發現原先的資料有誤，陸續補上正確資料），或是有各種例外格式需反覆調整擷取與轉換規則，並重新執行資料載入動作。此項儲藏資料的可重製性，可節省資料重整時間、提昇工作效率，以及達到隨時增進資料完整度與正確度的效益。

五、「用」為「使用」、「運用」之意，即建立完整的資料蒐集、整合與運用系統。

系統終端使用者不會在乎上述採、編、典、藏的流程和技術細節，一套符合需求的系統才是使用者關注的重點。根據不同的使用者類別，其需求可分述如下：

甲、資料輸入或提供者：系統的設計需特別注意資料蒐集介面與流程的簡化，例如：允許單一簽入、批次匯入、客觀性資料自動蒐整呈現免輸入、個人資料可隨時、隨地上網填報、存檔、列印、打包匯出或備份下載。

乙、一般行政人員：能接觸資料細節，對資料的分析需要詳細的觀點、友善的使用介面，可進行各種資料聚合（aggregation）與下探（drill down）操作，製作動態資訊儀表板，並產生日常的固定報表（Milam, Porter, & Rome, 2012）。

丙、行政與學術二級單位主管：可透過動態資訊儀表板，查看立即性動態資料與長期趨勢，協助研擬行政措施與策略。

丁、行政與學術一級單位主管：透過決策支援資訊系統，查看績效數據，進行風險管理。

戊、校務研究與教育數據分析人員：透過各種資料分析、統計、探勘技術與工具（Luan, Kumar, Sujitparapitaya, & Bohannon, 2012），進行特定任務分析報告，並建置校務研究分析作業流程，提供相關議題的諮詢以及支援各種分析任務。

己、各系所教研人員的專長研究：透過整合的倉儲資料，進行開創性議題的探索與研究。

上述關於資料蒐集、處理與運用的相關知識、技能、認知與態度，可稱是一種資料素養（Data Literacy）[5]。提升校務行政人員的資料素養，甚至教師的資料素養（Data Quality Campaign, 2014），對校務研究與學生的學習成效，將有很大的助益。

參、數據分析平台的建置方案

如前所述，數據分析平台在軟、硬體方面，近幾年有大幅的進展。Henschen（2014）曾比較 16 種各廠家與各類型的平台，底下將其歸納成三種建構教育數據分析平台的方案，並指出其優缺點。

一、傳統資料庫平台

倘若學校規模小，或是整體的校務資料庫所儲存與處理的資料量不大，仍可運用既有的資料庫管理系統，將各種較小規模的資料庫整合到軟、硬體規模較大的資料庫，並運用既有的統計軟體如 SPSS、SAS 或免費統計軟體 R（紀馥安、許清芳，2015），以及開源碼視覺化軟體[6]，自行組合出所需的系統功能。

其優點是無須投入太多軟、硬體成本，且既有的技術與方法可以沿用，缺點是各項資料分析功能與視覺化介面都要自行開發或整合，校務研究的系統建構較花人力與時間。

---

[5] 有關資料素養的多種定義，可參考：https://en.wikipedia.org/wiki/Data_literacy；跟教育有關的定義與說明，可參考：http://ites.ncdpi.wikispaces.net/Data+Literacy。

[6] 資料視覺化工具近幾年來發展迅速，簡介這些工具的網站也越來越多，相關網站可參見：20 Free Data Visualization Tools，https://codegeekz.com/free-data-visualization-tools/，存取日期 2015 年 6 月 20 日。

此方案在文獻上看到的案例，多為較早期或剛開始進行校務研究時，或規模小的機構所採用的方案。如 Schoenecker（2010）描述明尼蘇達州大學系統於 1995 年的發展情況，以及 Jones（2015）描述 Weber State University 於 1997 年購買一部伺服器開始進行校務研究的採用方案。

二、大數據資料庫平台

此部分以開源碼 Hadoop 平台為代表。Hadoop 是基於 Google 檔案系統（Ghemawat, Gobioff, & Leung, 2003）所開發的分散式大數據處理架構，其平台上各種工具軟體如圖 4 所示，包含：支援 SQL 查詢語言、ETL 工具、分散式資料庫、資料處理腳本語言（scripting language）、任務監控與管理工具等[7]。Hadoop 平台可以處理到 Petabytes（1000 的 5 次方，即 10 億兆位元組）等級的資料量或計算量，是傳統方案無法處理時的唯一選擇。

其優點是能處理極大量、成長速度極快的資料，且其為開源碼，軟體獲取的成本低。缺點是相關人員需要學習正在演進的各種平台工具與觀念，技術門檻高。

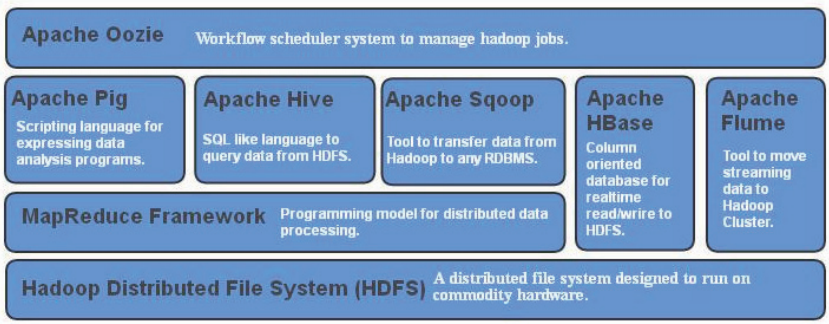


圖 4　Hadoop 平台各種工具軟體彼此之間的關係（Hadoop Ecosystem）[8]

文獻上看到的相關案例，有 Rios-Aguilar（2015）描述的案例研究，但不全然用到 Hadoop 這類架構。他們透過類似 Facebook 的行動軟體 Schools App，於 2011 年起在美國各地的九所社區學院邀請約 39,000 名教

[7] 更多相關工具軟體與計畫，詳見：http://hadoop.apache.org/，存取日期 2015 年 6 月 21 日。
[8] 圖片來源：http://blog.spec-india.com/apache-hadoop-an-introduction，存取日期 2015 年 6 月 15 日。

職員生共同參與，以蒐集其社交互動的資訊類型，並結合校務資料庫以瞭解學生成績與在校續讀率。他們花了近一年時間整合這些大量互動的各類型資料以及各種內容與格式的校務資料，並學習多種處理大數據的方法，包括檔案轉換（從 CSV 轉到 Excel、NVivo、UCINET、R、STATA 等）、分析方法（如：資料探勘、社會網路分析、經濟計量分析、質性分析等）、以及呈現方式（如：Word Clouds 詞雲、Sociograms 社會關係圖等），從而瞭解弱勢學生的各種狀況。

三、資料倉儲平台

另一種介於上述兩種極端的方案，是資料倉儲平台。其主要功能，是將組織中歷年累積下來的大量資料，透過其特殊的資料儲存架構（如：反正規化資料表、欄位為主的儲存方式、資料壓縮、大量主記憶體內的運算），便利資料分析的運用，如線上分析處理（On-Line Analytical Processing, OLAP）、資料探勘（Data Mining, DM）、資訊視覺化（Information Visualization, IV），進而協助建立動態資訊儀表板、決策支援系統等，以利即時回應外在環境變動，建構商業智慧（Business Intelligence, BI）。 因此，資料倉儲可說是 1970 年現代化的關聯式資料庫（Relational Database）理論被發展出來後，更進一步的資料儲存、整合與處理的概念與技術。

採用此方案的優點是能延續既有的資料庫技術，需要新學的技能較少，導入門檻較低，而又能獲得極佳的資料運算與分析效能，建立校務研究的基礎。缺點是目前的相關產品價格高、初期的建置成本與維護費用大。

美國多數大學皆採用此種方案，甚至組成高等教育資料倉儲論壇[9]，自 2005 年起，每年舉辦實務研討會議，交流彼此經驗。Schoenecker（2010）描述明尼蘇達州大學系統的案例，在 2010 年時，已整合 37 所機構的校務資料庫以及相關的校外資料庫，並引進資料倉儲、商業智慧、資料完整性等技術與概念，做為資料儲存、分析與視覺化的基礎。過程中他們分別雇用了商業智慧與資料庫設計公司以定義系統需求、設計資料架構、協助資料處理，並發展分析與報表軟體，同時也與大學系統內的資訊技術人員共同發展多面向的分析工具，以供及時的運用。

[9] The Higher Education Data Warehousing Forum，http://hedw.org/，存取日期 2015 年 7 月 11 日。

肆、校務研究資料庫的分析案例

校務研究的分析議題非常廣泛，從入學管理、生源分析、修退學原因、學習成效（彭森明，2010）、教育品質（羅孟彥，2013）、財務負擔、學術產出、教職員升遷等都有相關研究。從美國校務研究學會（Association for Institutional Research, AIR）推薦的期刊 New Directions for Institutional Research（NDIR）與 Research in Higher Education（RHE），以及日本近年來舉行的 Data Science and Institutional Research 研討會，可看出相關議題。底下以三項典型案例，說明在圖 1 的架構下，看到的資料素養問題，以及現行商用系統可立刻支援，或需要自行開發才能做到的分析與結果。

一、校務資訊統計年報

依據政府資訊公開法、大學法等規定，大學之校務資訊應公開，以保障人民知的權利。因此，各校於近年來，陸續以統計年報的方式，將校務資訊公開於各校網站上。以國立臺灣師範大學為例，統計年報的資料彙整與呈現範圍由各單位就其業務相關的資料進行填報，資料庫與網站系統則由資訊中心開發。此年報系統[10]，內容涵蓋：教務、學務、總務、研發、國際、師資培育以及畢業生就業等資料，是初步瞭解校務概況的重要資料庫。此系統建置於 2011 年，以當時技術其呈現方式，大致如圖 5 所示，每一網頁只用表格呈現資料數據，以及對應的圖片以視覺化這些資料。此系統的資料只能逐頁逐年呈現，無法將各年度、各學位的學生人數合併呈現出來，因而缺乏洞察趨勢與樣態的功能。

2013 年度學士班在學生人數

學院	人數	資料產生日期
教育學院	1760	2014-11-04 19:12:06
文學院	1540	2014-11-04 19:12:27
理學院	1562	2014-11-04 19:13:46
藝術學院	323	2014-11-04 19:15:03
科技學院	1082	2014-11-04 19:24:05
運動與休閒學院	497	2014-11-04 19:14:11
國際與僑教學院	550	2014-11-05 19:51:21
音樂學院	266	2014-11-05 19:52:07
管理學院	206	2014-11-04 19:14:42

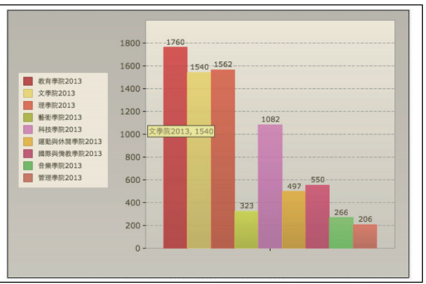


圖 5　臺師大統計年報資料呈現示意圖

[10] http://assess.ntnu.edu.tw/yreport/report_list.php，存取日期 2015 年 7 月 11 日。

為了補強其視覺化分析功能，並試驗前面兩節提到的概念與方案，我們將其中的年度、學院、學位、在學人數四個欄位資料從多個網頁中擷取、轉換到圖 1 的資料倉儲中，再利用視覺化分析工具，呈現如圖 6 的結果。

近年來有免費的視覺化軟體，如 Tableau Public[11]、Visualize Free[12]，標榜不用寫程式，只需連結資料、熟悉其視窗操作介面，即可呈現視覺化互動畫面，並發佈到雲端網頁上，供使用者在各種平台（包含平板、手機）上進行資料探索。圖 6 便是以 Tableau Public 連結到上述資料後，透過其視窗操作介面，整合年度、學院與學位的學生數，製作出來的視覺化結果。

從圖 6 中可看出幾項重點是原來系統中不易觀察出來的：一、從文學院、教育學院的年度序列數據，可明顯看出碩士生人數自 2011 年以後，與前面幾年的人數有明顯落差，如圖中圈選處。原因可能是對碩士生的定義不一致（例如：不列入碩士在職專班人數），或是受到少子化、就業市場的影響所致。若是後者，意味著各系所、學院與行政單位需要檢討並提出因應對策，以確保學校既有教學與行政人員的規模，或思考這是否導致較佳的師生比，對提升教學成效有所助益。二、科技學院 2014 年的學士生人數沒有出現，如圖中圈選處。原先以為是資料擷取、轉換與載入的過程有誤，查看原始網頁後發現，實際上該年沒有科技學院的資料，研判應該是資料輸入時有所疏漏。另外也發現學院名稱有不一致的情形，例如，在「國際與僑教學院」博士生人數各年度的網頁上，其學院名稱被登載成「國僑學院」；而在「運動與休閒學院」碩、博士生人數各年度的網頁上，其學院名稱則被登載成「運休學院」，兩者皆與該學院學士生的學院名稱不一致。這些不一致的情況，在圖 6 中（從下面的捲動軸往右移，可看到所有學院、學位與各年度的學生人數）即可快速發現。三、可看出某些學院的特殊情況，如社會科學學院沒有 2008 年的數據，是因為該學院於 2009 年才開始招收碩、博士生。而國際與僑教學院於 2014 年與社會科學學院合併成國際與社會科學學院。這些都是臺師大的特殊情況，但若因為資料輸入有誤或輸入不一致，對大眾而言，將會造成解讀上的困擾，無法確定能否相信這些特殊情況。

[11] https://public.tableau.com/s/，存取日期 2015 年 6 月 20 日。

[12] http://visualizefree.com/index.jsp，存取日期 2015 年 6 月 20 日。

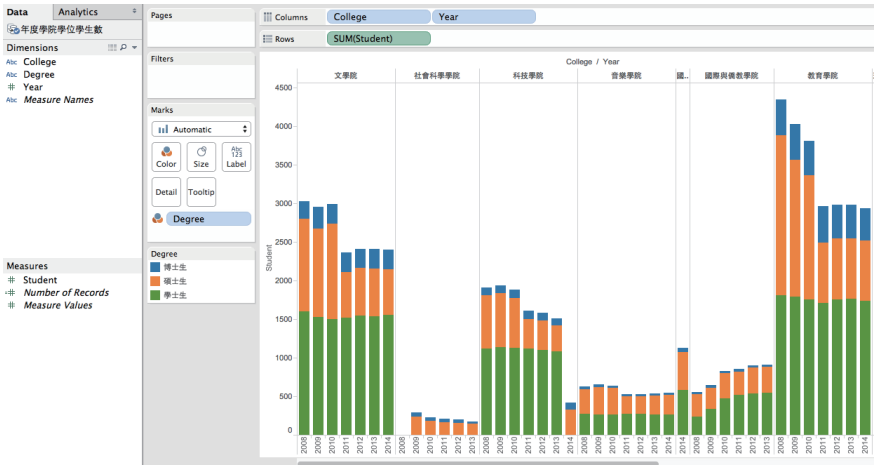


圖 6　以 Tableau Public 製作出來的資料視覺化結果

對應到前述的資料素養、處理流程，此案例顯示資料在採訪、典藏過程中，難以避免填報人員的命名不一致、資料缺漏的問題。然而透過資料整合以及新式視覺化軟體，可以容易偵測這些資料缺漏、驗證其正確性，進而通知填報單位改進，甚至找出值得探索的議題，讓校方探討其利弊得失。

二、Moodle 存取紀錄分析

大學的主要功能之一，是把教學做好，並充分利用持續演進的資通訊技術，以提升學習品質。在數位學習時代，各種學習平台、工具、軟體紛紛出現，這些學習輔助系統是否對學生有幫助，是值得研究的課題。特別是像 MOOCs（如：Coursera）、LMS（如：Moodle），以及 CRS（如：CloudClassRoom）等，可以蒐集到大量的使用者存取或回應紀錄，從而解析出提升學習成效的關鍵因子，例如透過 Moodle 瞭解學生的學習風格，以建議老師較佳的教學方案等（Graf, Kinshuk, & Liu, 2009）。

為此，我們運用臺師大 Moodle 系統一年約 2000 萬筆的存取紀錄，在數據分析平台採購流程的商情蒐集（Request for Information, RFI）階段，邀請廠商進行初步的概念驗證（Proof of Concept, POC），題目如附錄所

示，一方面瞭解各平台方案的優缺點、一方面可得知廠商的技術支援能力（廠商支援能力不夠，為 Jones（2015）曾述及的失敗經驗）。

結果，大部分的數據分析平台廠商在第一題：「將使用者依照存取 Moodle 的次數歸類成三到五類後（如：極少、偶而、經常），與使用的各項功能進行交叉統計」就做不出來。雖然歸類是統計、分析軟體的必備功能，但根據存取次數將使用者歸屬於不同類別，則需先統計每位使用者的存取次數，有基本統計數據後，設定存取次數門檻將使用者分類，才能與各欄位項目進行交叉分析，若無法看出有意義的樣態，則需再重設存取次數門檻，重複前述分析流程，如此反覆探索，直到找出（或無法找出）有意義的樣態為止。亦即在存取紀錄資料表中，除了要反正規化知道每位使用者的學號，並連結教務資料庫從學號知道其學院、系所、學位、年級外，還必須透過計算以增加一個欄位用來標示每位學生的使用頻率類別，才能進行概念驗證題目要求的各種交叉分析，繼而結合教務資料的修課成績以及學務資料的課外成就（如參加校內外比賽獲獎等），而找出影響學習成效的因子。

上述的概念驗證發現，廠商做不出第一題的情況有：不會對使用者依存取頻率分類（不知道要新增一個存取頻率類別的欄位），或不知道上述的反覆探索作法，或是低估反正規化所需的計算資源，甚至未預料到 Moodle 轉出的資料匯不進自己的系統。

然而第一題是泛用性的，亦即不管是 Moodle、MOOCs，還是其他的數位學習系統，只要使用者不來用，就沒有效果可以探討；使用者來用之後，可依其使用狀況將學生分類，結合其他屬性，如：所屬學院、家庭社經狀況、成績或成就等，以便從眾多的大量存取紀錄看出樣態，進行各種關聯探索與成效分析。就實務面而言，第一題需串連五個資料表，反正規化後才能得知學生學號及使用的功能。從學號再串接臺師大教務系統另外五個資料表，才能得知所屬院系、學位等資料，進行第二、三題的分析。這麼多資料表的串接，需依賴前述編、典、藏的確實作為，才得以釐清資料之間的複雜關係。

過去常看到以問卷方式，調查用戶使用某項服務或系統的頻率，用以交叉分析其使用成效。問卷中預設了用戶多久用幾次，就表示常用或少用

等類別，這常讓用戶需努力回想，再主觀或隨意的填答該項問題。然而線上系統有用戶的存取紀錄，善加利用，可以客觀的探索用戶使用樣態，較能降低使用成效分析的偏誤。在此案例中，將主觀意見量表，以客觀的既有數據取代，其關鍵流程在資料的編、典、藏作業以及擷取、轉換、載入（ETL）部分，而關鍵技術，則是具備多個資料表的反正規化能力，讓有些欄位無需用戶填答、再次蒐集，而是透過運算，提供應用。

三、教育學門國際論文分析

學校的聲望，是學生就學時選擇的參考項目，也是學生畢業後就業時被雇用的參考選項，更是學校能否吸引一流教研人員的重要因素之一。將學校聲望量化的方法之一，為大學之間的排名。因此，不論是世界性的、區域性的、領域性的排名，學校都要關注。

英國高等教育調查機構 QS 自 2011 年起公布世界大學排名，其 2015 年 4 月公布的「2014 年全球大學學科領域排名」，臺師大有 7 個學門入榜，其中教育學門全球第 22 名，為歷來最佳，國內排名第一，甚至超越許多國際知名大學。

為深入瞭解臺師大在教育學門的表現，特別是教育學門中各領域的排名，需用到校外的引文索引資料庫以及客觀的學門領域歸類分析，如 Diem 與 Wolter（2013）等人的研究。上述各廠商的數據分析平台都沒有現成的工具可用，因此我們自行發展了一套專為學術研究的內容分析工具 CATAR（Content Analysis Toolkit for Academic Research）（Tseng & Tsay, 2013），並放置於網路上提供各界免費下載[13]，以利運用於其他學門的分析，或是驗證我們發佈的結果。

我們於 2015 年 6 月 30 日從校外資料庫 Web of Science 下載教育領域 224 種期刊 2005 到 2014 十年間的論文共 66,740 篇到數據分析平台，根據其引用的 1,358,439 筆不同文獻，以期刊書目對（Journal Bibliographic Coupling）的相似度計算方式（Small & Koenig, 1977），用聚合式階層分群法（agglomerative hierarchical clustering）將 224 種期刊自動歸類成 36 個領域，其中數位學習與科學教育兩個領域的期刊歸類情況如圖 7 左邊所

[13] http://web.ntnu.edu.tw/~samtseng/CATAR/，存取日期 2015 年 6 月 20 日。

示。我們同時也用多維縮放技術（Multi-Dimensional Scaling, MDS）（Kruskal, 1997）將期刊相似度視覺化後顯示在二維空間中，如圖 7 右邊所示，其中主題近似的期刊在圖中距離也比較靠近。從圖 7 中可看出左上角分佈著科學教育領域的期刊，左下角為數位學習領域的期刊，從兩個領域的期刊名稱中，可看出利用期刊書目對的自動歸類結果相當合理。

表 2 列出數位學習與科學教育兩個領域論文發表數量前十名的大學。在這十年間，臺師大於數位學習領域論文數量排名全球第三，而國內共有七所大學進入前十名，顯示臺灣在此領域的學術研究相當活躍。在科學教育領域，國內則只有臺師大進入前十名，排名全球第二。此外，這十年間臺師大的論文總數有 316 篇，除了前述兩個領域佔了 154 + 70 = 224 篇外，在外語學習、高等教育、師資培育、學習科學等領域共 92 篇，也具有一定的國際能見度。

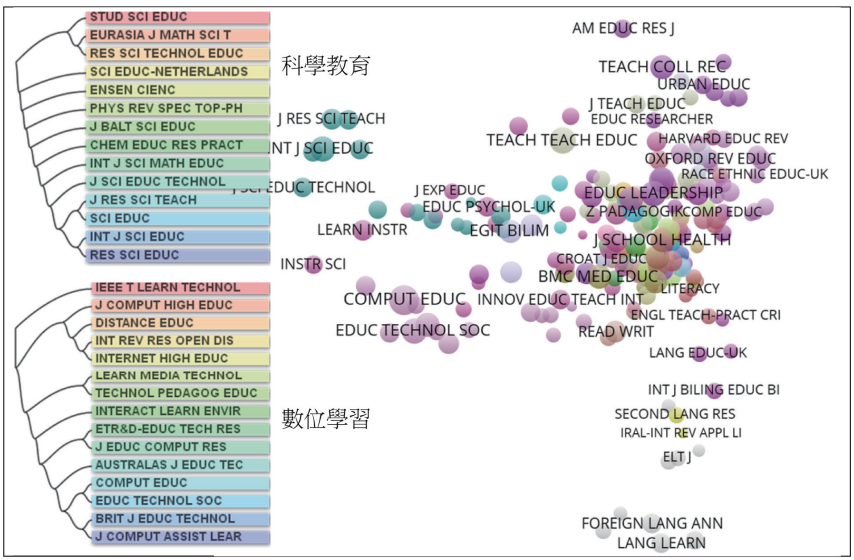


圖 7 教育學門 224 種期刊歸類後的視覺化結果

表 2　2005 ~ 2014 年數位學習與科學教育的機構排名

「數位學習」領域		「科學教育」領域	
機構名稱（縮寫）	論文數	機構名稱（縮寫）	論文數
Nanyang Technol Univ	180	Weizmann Inst Sci	82
Natl Cent Univ	158	Natl Taiwan Normal Univ	70
Natl Taiwan Normal Univ	154	Univ Colorado	65
Open Univ	149	Nanyang Technol Univ	64
Natl Taiwan Univ Sci & Technol	147	Middle E Technol Univ	60
Natl Cheng Kung Univ	103	Univ Missouri	59
Athabasca Univ	101	Curtin Univ Technol	57
Natl Chiao Tung Univ	79	Univ Michigan	55
Natl Sun Yat Sen Univ	75	Indiana Univ	55
Natl Univ Tainan	71	Michigan State Univ	54

透過上述的分析，可看出臺師大在教育各領域的國際學術成果，從而制訂學校研究發展的策略方向，培養具備潛力的頂尖研究團隊，更有效率地提升學校國際聲望。

此案例為臺師大每年進行的校務研究個案，已執行五年，逐步提升校方以數據決策的意識，據此所擬定的應對策略，已發揮成效，逐年提升臺師大教育領域的國際研究成果。

伍、結語

校務研究數據資料庫的建構與分析應用，是一項跨領域的議題：需具備教育與行政專業，以提出值得探討的問題，並有效解讀分析的結果；熟習圖書資訊學的知識，具備一定的資料素養，以建立可持久運作的資料處理作業流程；有資訊工程的技術，可串接各資料庫的欄位，反正規化成可提高分析效率的資料形式，以解決複雜的分析問題。當然，其他領域的知識與經驗，也會對此跨領域議題有所幫助。

本文介紹採、編、典、藏、用五項資料庫建構的流程與注意事項，提及資料素養的重要性，說明資料正規化與反正規化的用處，分析比較各種建置方案的適用時機與優缺點，透過概念驗證作業提供與國內廠商互動的經驗，並以實例分享分析與應用成果。

整體而言，資料蒐整作業（即：採、編、典、藏）仍是最費時、費力的流程，特別是串接各資料庫欄位，一旦完成，後續的分析運用便容易進行。然而真正有用的分析與決策輔助系統，通常難以事先規劃、建置完備。日常報告，自然有業務承辦人進行例行的處理，而需要決策時所須用到的資料，經常是非典型、外部的、未曾預見的。因此，除了本文介紹的知識與經驗，得從「做中學」累積實務歷練，這也是校務研究議題值得一再探究、分享彼此經驗之處。

目前的視覺化分析工具將越來越便利，讓各類型使用者得以更有效率的從大量資料中發現特殊樣態、形成假說，進而對資料做各種查詢與探索，以獲得具體事證支持決策。展望未來，視覺化工具不僅越來越受依賴，我們預期「事件演進模擬技術」（Bahr, 2009; Law & Kelton, 2000），將會是下一個數據分析平台的重要發展，其可讓使用者事先知道各種因素變化後的最終結果，讓分析平台更具威力。

誌謝

本文感謝教育部「邁向頂尖大學計畫」與科技部「跨國頂尖研究中心計畫」（MOST 104-2911-I-003-301）以及國立臺灣師範大學「華語文與科技研究中心」與「校務研究辦公室」贊助。另，感謝張國恩校長的支持，彭森明教授與王麗雲主任的諸多啟發，以及多位審查委員的建議，使文章內容前後呼應，特此申謝。惟文中所提論點純屬個人意見，不代表上述單位與人士之立場。

參考文獻

李欣宜（2015 年 2 月 17 日）。美國 Top 4 技術長寶立明：大數據即將在五年內消失。**數位時代**。取自 http://www.bnext.com.tw/。

[Li, X. Y. (2015, February 17). United States top 4 chief technology officer Stephen Brobst: Big data is about to disappear within five years. *Business Next.* Retrieved from http://www.bnext.com.tw/]

紀馥安、許清芳（2015）。運用開放軟體 R 處理大型教育資料庫。**當代教育研究季刊**，**23**(4)，121-153。

[Chi, F. A., & Sheu, C. F. (2015). Using R to analyze international large-scale educational assessment data. *Contemporary Educational Research Quarterly*, *23*(4), 121-153.]

彭森明（2010）。**大學生學習成果評量：理論、實務與應用。**臺北市：高等教育。

[Peng, S. S. (2010). *Accessing college student learning outcomes: Theory, practices, and applications*. Taipei, Taiwan: Higher Education.]

彭森明（2013）。**高等教育校務研究的理念與應用。**臺北市：高等教育。

[Peng, S. S. (2013). *Institutional research in higher education: Concepts and applications*. Taipei, Taiwan: Higher Education.]

曾元顯（2014）。自動化資訊組織與主題分析近二十年來的研究與發展。**教育資料與圖書館學**，**51**(5)，3-26。doi:10.6120/JoEMLS.2014.51S/0652.RV.AM

[Tseng, Y. H. (2014). Research and development on automatic information organization and subject analysis in recent decades. *Journal of Educational Media & Library Sciences, 51*(5), 3-26. doi:10.6120/JoEMLS.2014.51S/0652.RV.AM]

詹文男、羅瑋君（2001）。**高等資料庫報告－資料倉儲。**取自 http://www.mgt.ncu.edu.tw/~ylchen/database/DataWarehousing.doc

[Chan, W. N., & Lo, W. C. (2001). *Advanced database report－Data warehousing*. Retrieved from http://www.mgt.ncu.edu.tw/~ylchen/database/DataWarehousing.doc]

鍾沛原、曾賢寶、楊嘉麗、李柏毅、蔡一郎（2014）。**電腦機房異地備援機制參考指引。**取自 http://download.icst.org.tw/attachfilecomm/我國電腦機房異地備援機制參考指引.pdf

[Chung, P. Y., Tseng, H. B., Yang, J. L., Lee, B. Y., & Tsai, Y. L. (2014). *Reference guide to remote backup for computer & data center*. Retrieved from http://download.icst.org.tw/attachfilecomm/%E6%88%91%E5%9C%8B%E9%9B%BB%E8%85%A6%E6%A9%9F%E6%88%BF%E7%95%B0%E5%9C%B0%E5%82%99%E6%8F%B4%E6%A9%9F%E5%88%B6%E5%8F%83%E8%80%83%E6%8C%87%E5%BC%95.pdf]

羅孟彥（2013）。支援教育品質管理之整合式資訊系統建構與研究。**教育科學研究期刊**，**58**(4)，69-101。doi:10.6209/JORIES.2013.58(4).03

[Luo, M. Y. (2013). Design, implementation and study of an integrated information system for educational quality management. *Journal of Research in Education Sciences, 58*(4), 69-101. doi:10.6209/JORIES.2013.58(4).03]

Bahr, P. R. (2009). Educational attainment as process: Using hierarchical discrete-time event history analysis to model rate of progress. *Research in Higher Education*, *50*(7), 691-714. doi:10.1007/s11162-009-9135-x

Borthakur, D. (2007). The hadoop distributed file system: Architecture and design. *Hadoop Project Website*. Retrieved from https://svn.apache.org/repos/asf/hadoop/common/tags/release-0.16.4/docs/hdfs_design.pdf

Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, *13*(6), 377-387. doi:10.1145/362384.362685

Data Quality Campaign. (2014). *Teacher data literacy: It's about time—A brief for state policymakers.* Retrieved from http://dataqualitycampaign.org/wp-content/uploads/2015/06/DQC-Data-Literacy-Brief.pdf

Dey, E. L. (1997). Working with low survey response rates: The efficacy of weighting adjustments. *Research in Higher Education*, *38*(2), 215-227. doi:10.1023/A:1024985704202

Diem, A., & Wolter, S. C. (2013). The use of bibliometrics to measure research performance in education sciences. *Research in Higher Education, 54*(1), 86-114. doi:10.1007/s11162-012-9264-5

Feldman, R., & Sanger, J. (2006). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge, England: Cambridge University Press.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association, 64*(328), 1183-1210. doi: 10.1080/01621459.1969.10501049

Ghemawat, S., Gobioff, H., & Leung, S.-T. (2003). The Google file system. *ACM SIGOPS Operating Systems Review*, *37*(5), 29-43.

Graf, S., Kinshuk, & Liu, T.-C. (2009). Supporting teachers in identifying students' learning styles in learning management systems: An automatic student modelling approach modelling approach. *Educational Technology & Society*, *12*(4), 3-14.

Han, J., Kamber, M., & Pei, J. (2011). *Data mining: Concepts and techniques*. Burlington, MA: Morgan Kaufmann.

Henschen, D. (2014, January 30). 16 top big data analytics platforms. *InformationWeek*. Retrieved from http://www.informationweek.com

Hossler, D., Kuh, G., & Olsen, D. (2001). Finding fruit on the vines: Using higher education research and institutional research to guide institutional policies and strategies. *Research in Higher Education, 42*(2), 211-221. doi:10.1023/A:1026577604180

Inmon, W. H. (1992). *Building the data warehouse*. Hoboken, NJ: John Wiley & Sons.

Jin, L., Li, C., & Mehrotra, S. (2003). Efficient record linkage in large data sets. *Proceedings of the Eighth International Conference on Database Systems for Advanced Applications*, 137-146. doi: 10.1109/DASFAA.2003.1192377

Jones, L. (2015). How to build a data warehouse [The Higher Education Data Warehousing Forum]. Retrieved from http://hedw.org/hedwpresentation/how-tobuild-a-data-warehouse/

Kruskal, J. B. (1997). Multidimensional scaling and other methods for discovering structure. In K. Enslein, A. Ralston, & H. S. Wilf (Eds.), *Statistical methods for digital computers* (pp. 296-339). New York, NY: Wiley.

Kuhn, D. (2013). Data dictionary fundamentals. In D. Kuhn (Ed.), *Pro oracle database 12c administration* (pp. 259-275). New York, NY: Apress.

Larose, D. T. (2014). *Discovering knowledge in data: An introduction to data mining* (2nd ed.). Hoboken, NJ: Wiley.

Law, A. M., & Kelton, W. D. (2000). *Simulation modeling and analysis* (3rd ed.). New York, NY: McGraw Hill.

Luan, J., Kumar, T., Sujitparapitaya, S., & Bohannon, T. (2012). Exploring and mining data. In R. D. Howard, G. W. McLaughlin, & W. E. Knight (Eds.), *The handbook of institutional research* (pp. 1030-1077). Hoboken, NJ: John Wiley & Sons.

Milam, J., Porter, J., & Rome, J. (2012). Business intelligence and analytics. In R. D. Howard, G. W. McLaughlin, & W. E. Knight (Eds.), *The handbook of institutional research* (pp. 941-983). Hoboken, NJ: John Wiley & Sons.

Müller, H., & Freytag, J.-C. (2003). *Problems, methods and challenges in comprehensive data cleansing.* Retrieved from http://www.dbis.informatik.hu-berlin.de/fileadmin/research/papers/techreports/2003-hub_ib_164-mueller.pdf

Narayan, R. (1988). *Data dictionary: Implementation, use, and maintenance*. Upper Saddle River, NJ: Prentice-Hall.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science, 5*(1), 14-37. doi:10.1287/orsc.5.1.14

Plaisant, C. (2004). The challenge of information visualization evaluation. *Proceedings of the Working Conference on Advanced Visual Interfaces*, 106-116. doi: 10.1145/989863.989880

Rice, J. A. (2006). *Mathematical statistics and data analysis* (3rd ed.). Boston, MA: Duxbury Press.

Rios-Aguilar, C. (2015). Using big (and critical) data to unmask inequities in community colleges. *New Directions for Institutional Research, 163*, 43-57. doi: 10.1002/ir.20085

Schoenecker, C. (2010). The benefits of a comprehensive, integrated, and granular data system for community and technical college institutional research. *New Directions for Institutional Research, 147*, 81-108.

Simone, S., Campbell, C. M., & Newhart, D. W. (2012). Measuring opinion and behavior. In R. D. Howard, G. W. McLaughlin, & W. E. Knight (Eds.), *The handbook of institutional research* (pp. 1078-1119). Hoboken, NJ: John Wiley & Sons.

Small, H. G., & Koenig, M. E. D. (1977). Journal clustering using a bibliographic coupling method. *Information Processing & Management, 13*(5), 277-288.

Tillett, B. B. (2004). Authority control: State of the art and new perspectives. *Cataloging & Classification Quarterly, 38*(3-4), 23-41. doi:10.1300/J104v38n03_ 04

Tseng, Y.-H., Lin, C.-J., & Lin, Y.-I. (2007). Text mining techniques for patent analysis. *Information Processing & Management, 43*(5), 1216-1247. doi:10.1016/j.ipm.2006.11.011

Tseng, Y.-H., & Tsay, M.-Y. (2013). Journal clustering of Library and Information Science for subfield delineation using the bibliometric analysis toolkit: CATAR. *Scientometrics, 95*(2), 503-528.

Ware, C. (2012). *Information visualization: Perception for design*. Burlington, MA: Morgan Kaufmann.

Zhang, H., Wang, Y., & Han, J. (2011). Middleware design for integrating relational database and NOSQL based on data dictionary. *Proceedings of the International Conference on Transportation, Mechanical, and Electrical Engineering*, 1469-1472. doi: 10.1109/TMEE.2011.6199485

附錄：概念驗證題目與範例資料

題目：

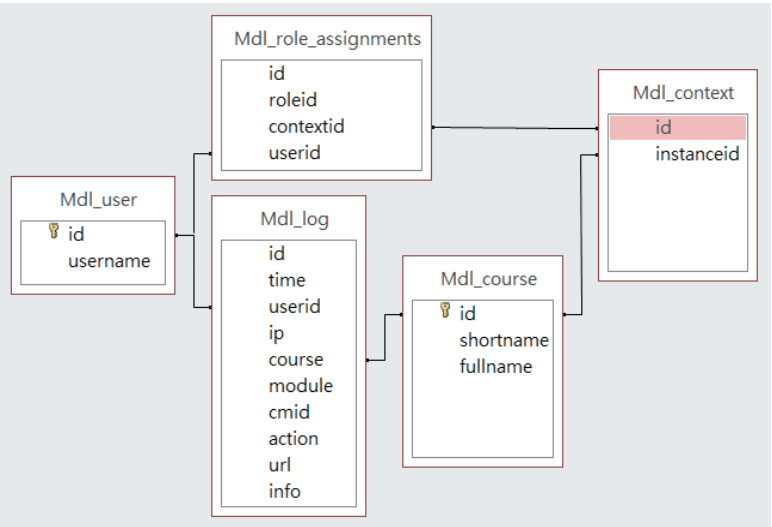
請就 Moodle 裡的存取記錄約 2000 萬筆，結合教務資料庫，分析：

一、將使用者依照存取 Moodle 的次數歸類成三到五類後（如：極少、偶而、經常），與使用的各項功能進行交叉統計，以瞭解 Moodle 的各項功能被使用的狀況。

二、各學院、各系所、甚至下探到（drill down）各課程的存取活動量比較表。

三、博士班、碩士班、大學部使用 Moodle 的每人平均次數（可再依每學期、每學院、或每門課分析）。

Moodle 資料庫：跟存取紀錄有關的資料表有五個，其關聯圖如下：



註：這些資料表的欄位與其之間的關係，是 Moodle 1.9 版的，這裡沒有將其修改、簡化，而選擇保留原來的關聯結構，除了可得知廠商對資料庫技術的深入程度，也考驗廠商系統反正規化處理的效率，更可提供採用 Moodle 系統的各個單位進行類似概念驗證時的直接運用。

教務資料庫：請直接以系統連接 Sybase 資料庫的方式，存取 STUDENT_VIEW 的資料（系統連接驗證資訊另外提供）。資料欄位及說明，請詳下表：

STUDENT_VIEW	
欄位名稱	說明
STD_NO	學號
SCH_SYS_M	學制代碼
SCH_CHIFULL	學制名稱（大學、碩、博）
COL_CODE	學院代碼
COL_CHIFULL	學院名稱
DEPT_CODE	系所代碼
DEPT_CHIFULL	系所名稱
DEPT_GROUP	組別代碼
DPT_CHIFULL	組別名稱
FORM_S	年級
STD_NO	學號

註： 由於教務系統龐大，且各校不盡相同，我們事先從五個資料表擷取跟題目相關的欄位，反正規化後供廠商進行概念驗證。